INFO 6210

Data Management and Database Design

Database Project Proposal

AI Skunkworks Project

Hyperparameter Database

Contents: -

- Project Description
  - Goals and objectives
  - Project Requirement
  - Problems to be addressed
  - Potential pitfalls & challenge
- Algorithms and code sources
- Data sources
- References
- Project members

Project Description

Goals and objectives:

Gather a list of data sets, type of datasets, and hyperparameters by running an expanded list of datasets. This information will be embedded in a database management system, to be incorporated into a website where it is easy to be searched and used by the public.

The hyperparameter database is created by running millions of hyperparameter values, over thousands of public datasets and calculating the individual conditional expectation of every hyperparameter on the quality of a model.

Generate models using H2O software to find the best hyperparameters and create a conceptual model and store all the data into a physical database.

Project Requirement:

Unique datasets are to be picked from different data sources like Kaggle Datasets, UCI machine learning repository, Amazon Datasets, Google Datasets, Computer Vision Datasets etc. Identify the type of dataset chosen, ie Regression, Classification, Clustering etc. Perform data cleaning and data pre-processing. Create conceptual and ER diagrams. Perform database normalization and perform analytics on the database created to get the best values for the hyperparameters.

Problems to be addressed:

Most of the algorithms that improve metrics, degrades the quality of search results. Hyperparameter optimization is performed to overcome the issues addressed by those algorithms and build models for visualizing and teaching statistical concepts.

Potential pitfalls & challenge:

Different optimization methods will have different setup steps, time requirements, and performance outcomes. Hence, methods like algorithmic optimization will help in achieving better performance.

Algorithms and code sources

The effect of hyperparameters is analyzed in the hyperparameter database using the following algorithms: Distributed Random Forest (DRF), Generalized Linear Model (GLM), Gradient Boosting Machine (GBM). Naïve Bayes Classifier, Stacked Ensembles, Xgboost and Deep Learning Models (Neural Networks).

Data sources

Data is derived from data sources like Kaggle Datasets, UCI machine learning repository, Amazon Datasets, Google Datasets, Computer Vision Datasets etc.

References

- https://en.wikipedia.org/wiki/Hyperparameter_(machine_learning)
- https://towardsdatascience.com/top-sources-for-machine-learning-datasets-bb6d0dc3378b
- https://en.wikipedia.org/wiki/List_of_datasets_for_machine-learning_research
- https://medium.com/@alexandraj777/top-5-mistakes-data-scientists-make-with-hyperparameter-optimization-and-how-to-prevent-them-767638b245f8
- https://stats.stackexchange.com/questions/297337/what-are-some-of-the-disavantage-of-bayesian-hyper-parameter-optimization
- https://github.com/skunkworksneu/Projects

Project members

1. Megha Ponneti Nanda [001388342]
2. Pratiksha Milind Lavhatre [001388250]